Set-based replay attack detection in closed-loop systems using a plug & play watermarking approach

Carlos Trapiello[1,2] and Vicenç Puig[1,2]

Abstract— **This paper presents a watermarking signal injection method that compensates its effect in the loop, avoiding thus the signal reinjection. Similar to a virtual actuator scheme, the proposed methodology masks the presence of the authentication signal to the system controller, that do not need to be retuned as it remains immunized. Furthermore, a set-based analysis concerning the effect that the performance loss imposed by a watermarking signal has in the detectability of a replay attack is performed for the stationary, assuming that a standard state observer is used in order to monitor the plant. Finally, a numerical application example is used to illustrate the proposed approach.**

I. INTRODUCTION

Cyber-physical systems (CPS) consist of both, logical and physical elements connected by communication channels. The advantages in efficiency and adaptability that CPSs entail have led to its successful implementation in different sectors like water and gas supply systems, smart grids or transportation networks among others. Nevertheless, their capability to integrate and coordinate heterogeneous components also imply the introduction of new security weaknesses that could be exploited by a malicious attacker[1]. Registered attacks like [2], [3] have demonstrated the severe consequences that planned attacks against critical infrastructures may imply to society.

In [4], malicious cyber attacks are classified as either *denial-of-service* (DoS) or *deception* attacks. While DoS attacks consist in blocking the communication between system components (sensors, actuators and/or controllers), deception attacks provoke that one, or several, system components receive false information and believe it as true, being therefore harder to detect. Among the different deception attacks considered in the literature [5], replay attacks appear as one of the most common and natural ones to be launched by an attacker who does not have any *a priori* knowledge of the system dynamics, but is aware that the system itself will be in steady state during the attack [6]. For the replay attack considered in this work, it is assumed that: 1) a malicious attacker secretly records a time slot of the sensing data sent to a monitoring station; 2) the recorded data is replayed back in order to mask a physical attack conducted over the plant.

Security in control systems is not a new topic of study, and the problem of fault diagnosis and fault tolerant control has been thoroughly studied in the literature. Despite the resemblances between the attack detection and fault diagnosis problems, replay attacks have shown to be undetectable by means of standard fault detection algorithms (in [6] the authors have shown the replay attack undetectability by some statistical detectors, like the χ^2 detector). From a consistency based point of view, fault diagnosis techniques analyze the given I/O pair (U, Y) together with a nominal plant behavior reference, in order to detect, isolate and estimate the fault [7]. However, when a replay attack substitute the actual measurements by previously recorded ones ($\bar{Y}$), data consistency is preserved, achieving to deceive fault detection schemes. As consequence, it seems necessary to modify the system inputs with the inclusion of an authentication signal (ΔU), in order to analyze the new sensor measurements (Y') checking for the time-dependent injected signal. Nevertheless, this comes at the cost of introducing a performance loss that separates the new system outputs (Y') from the nominal ones (Y).

The aforementioned detection techniques are placed within the watermarking-based approach. Some of the authentication signals proposed in the literature in order to detect replay attacks are: In [6], an independent and identically distributed (IID) Gaussian distribution is used to generate the authentication signal; the watermark proposed in [8] aims at destabilizing the residual while preserving the stability of the main system; in [9] the authors employ a periodic watermarking strategy; in [10], a sinusoidal signal with a time-varying frequency is proposed as possible signature; and in [11] a multi-agent extension of the watermarking concept is presented. It must be remarked the introduction of alternative methods (see [12]) which, under certain assumptions, propose replay attack detection techniques without injecting an exogenous signal in the control inputs.

The contribution of this paper is twofold: 1) The introduction of a novel methodology for injecting a watermarking signal such that, based on virtual actuator (VA) schemes, compensates its effect in the system outputs with the aim to avoid the reinjection in the loop, and thus immunizing the system controller that do not need to be retuned; 2) The development of a robust study using set-based techniques, with the purpose of characterizing for the stationary the relationship between the required system outputs performance loss, and the detectability of the replay attack if a classical

This work has been partially funded by the Spanish State Research Agency (AEI) and the European Regional Development Fund (ERFD) through the projects SCAV (ref. MINECO DPI2017-88403-R) and DEOCS (ref. MINECO DPI2016-76493) and AGAUR ACCIO RIS3CAT UTILITIES 4.0 – P7 SECUTIL. This work has been also supported by the AEI through the Maria de Maeztu Seal of Excellence to IRI (MDM-2016-0656).

[1] The authors are with the Research Center for Supervision, Safety and Automatic Control (CS2AC) of the Universitat Politècnica de Catalunya (UPC), Spain carlos.trapiello@upc.edu

[2] The authors are also with the Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain.

state observer is used in order to generate the residual set. This study is carried out under the assumption of closed-loop state feedback.

The remaining of the paper is structured as follows: Section II describes the proposed method for injecting a watermarking signal without affecting the controller. In Section III, a robust analysis of the performance loss induced by a watermarking signal is carried out using a set-based approach. Section IV presents a numerical example in order to demonstrate the validity of the proposed approach. Finally, the main conclusions are drawn in Section V.

II. Watermarking signal injection

In this section, a design mechanism for injecting a watermarking signal without affecting the predefined system controller is proposed.

A. Formulation

Let us consider a discrete-time linear time-invariant (LTI) system

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k \\ y_k &= Cx_k \end{aligned} \tag{1}$$

where $u \in \mathbb{R}^{n_u}$ is the input vector, $y \in \mathbb{R}^{n_y}$ the output vector, $x \in \mathbb{R}^{n_x}$ is the state vector, and A, B, C are the state-space matrices of appropriate dimensions. A square system is assumed in this work, i.e. $n_u = n_y$. The control objective is to regulate the plant tracking error

$$z = x - x_r \tag{2}$$

where the reference signal x_r satisfies the reference model[1]

$$\begin{aligned} x_r^+ &= Ax_r + Bu_r \\ y_r &= Cx_r \end{aligned} \tag{3}$$

The defined tracking error is stabilized by means of a control action of the form

$$u = f(z, y_r) \tag{4}$$

being y_r the imposed set-point.

In order to force the detection of a malicious attack affecting system (1), a watermarking signal Δu is added through the system inputs u. With this purpose, let us consider a new set of states x_α that depend on the plant outputs y as follows

$$x_\alpha^+ = y + B_\alpha d = Cx + B_\alpha d \tag{5}$$

where $d \in \mathbb{R}^{n_y}$ is an exogenous signal and B_α is an input matrix such that the pair (C, B_α) is controllable. Defining $x_2 = \begin{bmatrix} x & x_\alpha \end{bmatrix}^T$, systems (1) and (5) can be gathered together as

$$\begin{aligned} x_2^+ &= A_2 x_2 + B_2 u_2 \\ y &= C_2 x_2 \end{aligned} \tag{6}$$

with

$$u_2 = \begin{bmatrix} u \\ d \end{bmatrix} \; A_2 = \begin{bmatrix} A & 0 \\ C & 0 \end{bmatrix} \; B_2 = \begin{bmatrix} B & 0 \\ 0 & B_\alpha \end{bmatrix} \; C_2 = \begin{bmatrix} C & 0 \end{bmatrix} \tag{7}$$

Furthermore, let us consider another input matrix for system (5) with the form: $B_\alpha^\star = 0$, such that the control of the new plant can only be exerted by means of system (1) outputs. Under this consideration, the dynamics of the overall system (6) become

$$\begin{aligned} x_2^{+,\star} &= A_2 x_2^\star + B_2^\star u_2 \\ y &= C_2 x_2^\star \end{aligned} \tag{8}$$

where

$$x_2^\star = \begin{bmatrix} x \\ x_\alpha^\star \end{bmatrix} \quad B_2^\star = \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} \tag{9}$$

resulting in an under-actuated control system, i.e. the equilibrium points of the new system (5), are imposed by the equilibrium points of system (1). According to this limitation, a virtual actuator scheme [7] may be used for adding a signal Δu to the nominal control action (4) (yielding a new control signal $u' = u + \Delta u$), such that, alters the system (1) outputs for stabilizing the states x_α in a desired set point.

Note that, if the pair (A, B) is controllable and if $rank(C) = n_y$, then, the pairs (A_2, B_2) and $(A_2, B_2^\star)$ are controllable for any matrix B_α, as A_2 is only dependent on the system (1) states.

Similar to VA schemes, the imposed rank loss in the input matrix

$$rank(B_2^\star) < rank(B_2) \tag{10}$$

motivates the use of a new dynamics system, called difference system, in order to control the under-actuated system (8). This yields a new control action u_2'

$$u_2' = Nu_2 + Mx_\Delta \tag{11}$$

where the matrix N is

$$N = {B_2^\star}^\dagger B_2 \tag{12}$$

being ${B_2^\star}^\dagger$ the pseudoinverse of $B_2^\star$. The dynamics of the difference system x_Δ are given by

$$\begin{aligned} x_\Delta^+ &= (A_2 - B_2^\star M)x_\Delta + (B_2 - B_2^\star N)u_2 \\ y_\Delta &= C_2 x_\Delta \end{aligned} \tag{13}$$

with the matrix M designed such that $(A_2 - B_2^\star M)$ is a Schur matrix. The specific form of $B_2^\star$ implies that

$$N = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \; M = \begin{bmatrix} M' \\ 0 \end{bmatrix} \; B_\Delta = B_2 - B_2^\star N = \begin{bmatrix} 0 & 0 \\ 0 & B_\alpha \end{bmatrix} \tag{14}$$

Hence, the new control action (11) adopts the form

$$u_2' = \begin{bmatrix} u' \\ 0 \end{bmatrix} = \begin{bmatrix} u \\ 0 \end{bmatrix} + \begin{bmatrix} M' x_\Delta \\ 0 \end{bmatrix} = \begin{bmatrix} u + \Delta u \\ 0 \end{bmatrix} \tag{15}$$

i.e. the new control action u' injected to the original system (1), is now extended with the addition of $\Delta u = M' x_\Delta$.

[1] Henceforth, the index $k+1$ will be replaced by $+$ and k will be omitted for the sake of simplified notations

According to (14), the dynamics of the difference system (13) are only actuated by means of the externally imposed signal d as

$$\begin{aligned} x_\Delta^+ &= (A_2 - B_2^\star M)x_\Delta + B_\Delta u_2 = (A_2 - B_2^\star M)x_\Delta + \begin{bmatrix} 0 \\ B_\alpha \end{bmatrix} d \\ y_\Delta &= C_2 x_\Delta \end{aligned} \tag{16}$$

The whole reconfigured plant may be expressed as follows

$$\begin{aligned} \begin{bmatrix} x_2'^+ \\ x_\Delta^+ \end{bmatrix} &= \begin{bmatrix} A_2 & B_2^\star M \\ 0 & A_2 - B_2^\star M \end{bmatrix} \begin{bmatrix} x_2' \\ x_\Delta \end{bmatrix} + \begin{bmatrix} B_2^\star N \\ B_\Delta \end{bmatrix} u_2 \\ \begin{bmatrix} y' \\ y_\Delta \end{bmatrix} &= \begin{bmatrix} C_2 & 0 \\ 0 & C_2 \end{bmatrix} \begin{bmatrix} x_2' \\ x_\Delta \end{bmatrix} \end{aligned} \tag{17}$$

therefore, by adding the outputs y_Δ to the system outputs y' (see Fig. 1), the dynamics of the complete reconfigured system, as seen from the controller point of view, can be modeled (defining $\tilde{x} = x_2' + x_\Delta$) as

$$\tilde{x}^+ = A_2\tilde{x} + B_2 u_2 \tag{18a}$$

$$y = C_2\tilde{x} = C_2 x_2' + C_2 x_\Delta = y' + y_\Delta \tag{18b}$$

that has the same dynamics as system (6), and hence the dynamics of the original plant (1).

B. Injected signal

According to (18b), by imposing a watermarking signal d that affects the difference system outputs y_Δ, the same effect, but with opposite sign, will be produced in the plant outputs y' through the addition of the derived $\Delta u = M' x_\Delta$ control signal.

The advantage of injecting the authentication signal d following the proposed scheme is that the system controller will not notice the addition of the new control action Δu. Therefore, no controller retuning is needed in order to cope with the reinjection of the watermark signal in the control loop. This methodology can be perceived as a *plug & play* mechanism [13] allowing to include a watermarking signal in an already defined control loop.

Throughout the rest of the paper, the considered authentication signal will be a random switch in the system set-point y_r', by modifying its components between the values

$$y_{r,i}' \in \{y_{r,i} - \frac{\delta y_i}{2}\ ,\ y_{r,i} + \frac{\delta y_i}{2}\} \tag{19}$$

where $y_{r,i}$ represents the desired value in the i^{th} system output imposed by the nominal system controller (4), and δy_i represents the set-point offset forced by the authentication signal Δu. This is done in order to achieve attack detectability when a temporal mismatch between the expected and the measured output is produced as a consequence of an attack.

The imposed set-point offset, denoted as $\Delta y_r = [\delta y_1, \cdots, \delta y_{n_y}]^T$, is therefore obtained by injecting an external signal d such that forces the desired offset (with opposite sign) in the difference system outputs y_Δ.

Remark: The selection of the previous authentication signal is motivated in order to link with performance loss analysis developed in the following section. Nevertheless, the same procedure holds if more complex watermarking signals are injected, remaining the controller immunized in any case.

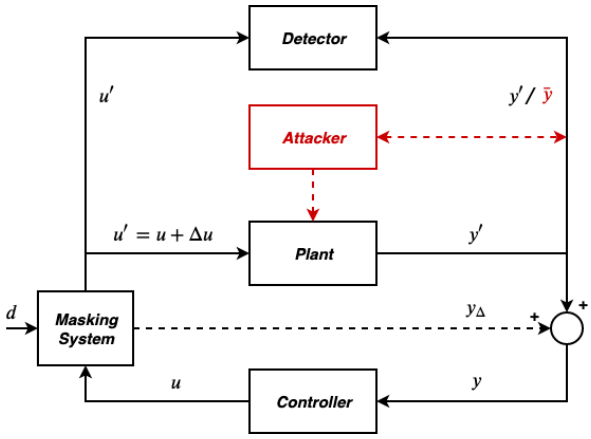

Fig. 1. Watermarking signal injection scheme

III. Set-based replay attack detection

In this section, set-based approaches (that has been successfully applied to the FDI problem [14]) are used in order to robustly analyze the performance of classical detectors, under the presence of a replay attack, as a function of the performance loss introduced by a watermarking signal.

This performance loss must be understood, element by element, as the set-point difference Δy_r that exist between the set reference when the signal was recorded by a malicious attacker $\bar{y}_{r,i}$, and the imposed output reference $y_{r,i}$ when the attack is being deployed, i.e. when the recordings $\bar{y}_i$ are replayed back by substituting the real sensor measurements y_i.

As shown in Fig. 1, the considered replay attack is assumed to affect only the measurements sent to an anomalies detector. Furthermore, the presence of a stabilizing low-level feedback control loop is also assumed throughout the present section.

A. Preliminaries

For the following development, the discrete-time LTI system (1) is extended by considering the presence of process disturbances w and measurement noise v, resulting in a system with the form

$$\begin{aligned} x^+ &= Ax + Bu + E_w w \\ y &= Cx + E_v v \end{aligned} \tag{20}$$

Moreover, both, process and measurements noise, are assumed to be unknown but bounded

$$\begin{aligned} \mathcal{W} &= \{w \in \mathbb{R}^{n_x} : |w - c_w| \leq \bar{w}, c_w \in \mathbb{R}^{n_x}, \bar{w} \in \mathbb{R}^{n_x}\} \\ \mathcal{V} &= \{v \in \mathbb{R}^{n_y} : |v - c_v| \leq \bar{v}, c_v \in \mathbb{R}^{n_y}, \bar{v} \in \mathbb{R}^{n_y}\} \end{aligned} \tag{21}$$

where c_w, $\bar{w}$, c_v and $\bar{v}$ are constant vectors, and the inequalities in (21) are considered component-wise.

The sets expressed in (21) can be rewritten using a zonotopic representation [15] as

$$\begin{aligned}\mathcal{W} &= \langle c_w, R_w \rangle \\ \mathcal{V} &= \langle c_v, R_v \rangle\end{aligned} \tag{22}$$

where c_w and c_v denote the centers of the disturbance and noise bounding zonotopes respectively, with their respective generator matrices $R_w \in \mathbb{R}^{n_x \times n_x}$ and $R_v \in \mathbb{R}^{n_y \times n_y}$. Hereinafter it will be assumed that disturbance and noise are bounded by a unitary hypercube zonotopes centered at the origin

$$\begin{aligned}\mathcal{W} &= \langle 0, I_{n_w} \rangle \\ \mathcal{V} &= \langle 0, I_{n_v} \rangle\end{aligned} \tag{23}$$

where $I_{n_w} \in \mathbb{R}^{n_w \times n_w}$ and $I_{n_v} \in \mathbb{R}^{n_v \times n_v}$ denote the identity matrices.

In order to regulate the tracking error with respect the reference model (3), a linear fixed gain feedback control action is considered

$$u = u_r + v = u_r - Kz \tag{24}$$

By taking into consideration (20) and (3), the tracking error dynamics are defined as

$$z^+ = Az + Bv + E_w w = (A - BK)z + E_w w \tag{25}$$

with K designed such that $(A-BK)$ is a Schur matrix. The dynamics (25) represent a stable LTI system with bounded additive disturbances (23). Thus, a zonotopic representation $\mathcal{Z}$ of the robust positive invariant (RPI) set for z can be defined as (see [16])

$$\mathcal{Z} = \langle c_{z_\infty}, R_{z_\infty} \rangle \tag{26}$$

where c_{z_∞} and R_{z_∞} are computed by recursively propagating the zonotopic set

$$\begin{aligned}c_{z_{j+1}} &= (A - BK)c_{z_j} \\ R_{z_{j+1}} &= [(A - BK)R_{z_j} \;\; E_w]\end{aligned} \tag{27}$$

starting from the initial zonotope $\mathcal{Z}_0 = \langle c_{z_0}, R_{z_0} \rangle$ that belongs to the RPI set which could be calculated by means of any of the developed procedures (e.g. Ultimate Bound [17]). The zonotopic representation (27) will converge to a smaller RPI set in the steady state, i.e., when $j \to \infty$.

B. Residual generator

The studied detector consists on the classical residual construction used in the FDI community

$$r = y - \hat{y} \tag{28}$$

where y represents the vector of measured outputs, subject to a replay attack, and $\hat{y} = C\hat{x}$ represents the output estimation.

The monitoring of the plant system (20) can be done by designing a Luenberger observer of the form

$$\hat{x}^+ = A\hat{x} + Bu + L(y - \hat{y}) \tag{29}$$

where $\hat{x}$ is the state estimation and the observer gain L is computed such that $(A - LC)$ is a Schur matrix.

Denoting $\tilde{x} = (x - \hat{x})$ as the state estimation error, then, from (20) and (29), the estimation error dynamics are expressed as

$$\tilde{x}^+ = (A - LC)\tilde{x} + \begin{bmatrix} E_w & -LE_v \end{bmatrix} \begin{bmatrix} w \\ v \end{bmatrix} = (A - LC)\tilde{x} + E_\eta \eta \tag{30}$$

Similar to the tracking error dynamics (25), the system (30) represents a stable LTI system with bounded additive disturbances (23). Therefore, a zonotopic representation $\tilde{\mathcal{X}}$ of the RPI set for the estimation error $\tilde{x}$ is denoted as

$$\tilde{\mathcal{X}} = \langle c_{\tilde{x}_\infty}, R_{\tilde{x}_\infty} \rangle \tag{31}$$

where $c_{\tilde{x}_\infty}$ and $R_{\tilde{x}_\infty}$ are computed as the stationary values, i.e. $j \to \infty$, of the propagating zonotopic set

$$\begin{aligned}c_{\tilde{x}_{j+1}} &= (A - LC)c_{\tilde{x}_j} + c_w - Lc_v \\ R_{\tilde{x}_{j+1}} &= [(A - LC)R_{\tilde{x}_j} \;\; E_w \;\; -LE_v]\end{aligned} \tag{32}$$

starting from the initial zonotope $\tilde{\mathcal{X}}_0 = \langle c_{\tilde{x}_0}, R_{\tilde{x}_0} \rangle$ belonging to a RPI set.

1) Healthy functioning: For the healthy functioning, i.e. when no attack is performed over the plant outputs, the residual set has the form

$$r^H = y - \hat{y} = Cx + E_v v - C\hat{x} = C\tilde{x} + E_v v \tag{33}$$

Thus, projecting the zonotopic representation of the RPI sets associated to the measurements noise (23) and the estimation error (31), into the healthy residual space, the zonotopic representation of a RPI healthy residual set is obtained as

$$\mathcal{R}^{\mathcal{H}} = \langle c_{r_H}, R_{r_H} \rangle \tag{34}$$

with

$$\begin{aligned}c_{r_H} &= Cc_{\tilde{x}_\infty} = 0 \\ R_{r_H} &= [CR_{\tilde{x}_\infty} \;\; E_v]\end{aligned} \tag{35}$$

Note that the size of the remaining residual set is dependent on the selection of the observer gain L.

2) System under replay attack functioning: Whenever a subset l ($l \subseteq m \,|\, m = \{1, ..., n_y\}$) of the system outputs is substituted by previous recordings, i.e. $\forall i \in l \Rightarrow y_i = \bar{y}_i$, yielding a new output vector $\bar{y}$, then, the monitoring observer (29) looses the stabilizing loop being its dynamics governed by two independent and bounded inputs $(u, \bar{y})$.

If the error between the system state $\bar{x}$ that would yield the substituted outputs $\bar{y}$ and the estimated state $\hat{x}$ is taken under consideration, then, the residual set may be expressed as

$$r^A = \bar{y} - \hat{y} = C\bar{x} + E_v v - C\hat{x} = Ct + E_v v \tag{36}$$

where $t = \bar{x} - \hat{x}$.

The dynamics of this new estimation error system are

$$\begin{aligned}t^+ &= \bar{x}^+ - \hat{x}^+ = A\bar{x} + B\bar{u} + E_w w - (A\hat{x} + Bu + L(\bar{y} - \hat{y})) \\ &= (A - LC)t + B(\bar{u} - u) + E_w w - LE_v v\end{aligned} \tag{37}$$

with

$$\bar{u} = \bar{u}_r - K\bar{z} \quad u = u_r - Kz \tag{38}$$

denoting $\Delta u_r = \bar{u}_r - u_r$, then

$$t^+ = (A-LC)t + B\Delta u_r - K\bar{z} + Kz + E_w w - LE_v v \tag{39}$$

The previous equation represents an LTI system with bounded additive disturbance, by using (23) and (26) the zonotopic representation of the RPI set for the previously defined estimation error during the attack may be expressed as

$$\mathcal{T} = \langle c_{t_\infty}, R_{t_\infty} \rangle \tag{40}$$

where c_{t_∞} and R_{t_∞} are computed as the stationary values, i.e. $j \to \infty$, of the following propagation zonotopic set

$$\begin{aligned} c_{t_{j+1}} &= (A - LC)c_{t_j} + B\Delta u_r \\ R_{t_{j+1}} &= [(A - LC)R_{t_j} \ -BKR_{z_\infty} \ BKR_{z_\infty} \ E_w \ -LE_v] \end{aligned} \tag{41}$$

starting from, the initial zonotope $\mathcal{T}_0 = \langle c_{t_0}, R_{t_0} \rangle$ belonging to a RPI set.

Projecting the previous zonotopic representation of the estimation error RPI set (40), in the attacked residual space defined in (36)

$$c_{r_A} = Cc_{t_\infty} = C(I - (A - LC))^{-1}B\Delta u_r \tag{42a}$$

$$R_{r_A} = [CR_{t_\infty} \ E_v] \tag{42b}$$

Note that, under the assumption that system matrix A has no integrating modes, equation can be used (3) to express the difference in the reference set-point $\Delta y_r = \bar{y}_r - y_r$ as

$$\Delta y_r = C(I - A)^{-1}B\Delta u_r \tag{43}$$

Considering the equality[2]

$$(I-(A-LC))^{-1} = (I-A)^{-1} - (I-(A-LC))^{-1}LC(I-A)^{-1} \tag{44}$$

the expression (42a) can be reformulated as

$$\begin{aligned} c_{r_A} &= C(I - (A - LC))^{-1}B\Delta u_r \\ &= C((I - A)^{-1} - (I - (A - LC))^{-1}LC(I - A)^{-1})B\Delta u_r \\ &= (I - C(I - (A - LC))^{-1}L)\Delta y_r \end{aligned} \tag{45}$$

Comparing the derived healthy and attack residual generator matrices, it can be seen that the size of the obtained RPI attack set is bigger that the healthy one. This conservatism is consequence of the independence between the injected control input u and the measured control output $\bar{y}$ during the attack. According to that, replay attack may be detected even without forcing a performance loss in the system.

Equation (45) shows how the difference between the set-point of the replayed signal $\bar{y}_r$ and the set-point imposed through the control action y_r, shifts the center of the attack residual set. The previous difference may be achieved by randomly switching the system reference between predefined values. Note that the center displacement of the RPI attack set, given an obtained Δy_r, follows some directions imposed by the system dynamics and the monitoring observer gain L.

[2]Equality derivation:

$$I - A = (I - (A - LC)) - LC$$

$$(I - (A - LC))^{-1}(I - A) = I - (I - (A - LC))^{-1}LC$$

$$(I - (A - LC))^{-1} = (I - A)^{-1} - (I - (A - LC))^{-1}LC(I - A)^{-1}$$

IV. Numerical example

Let us consider a discrete time LTI system, see (20), with the following system matrices

$$A = \begin{bmatrix} 0.9842 & 0.0407 \\ -0.1327 & 0.9590 \end{bmatrix} \quad B = \begin{bmatrix} 0.0831 & 0.0007 \\ 0 & 0.0352 \end{bmatrix}$$
$$C = \begin{bmatrix} 0.1000 & 0 \\ 0 & 0.0500 \end{bmatrix} \tag{46}$$

with disturbance and process noise distribution matrices as

$$E_w = \begin{bmatrix} 0.0500 & 0 \\ 0 & 0.0500 \end{bmatrix} \quad E_v = \begin{bmatrix} 0.0100 & 0 \\ 0 & 0.0100 \end{bmatrix} \tag{47}$$

and a sample time $T_s = 1s$.

The previous system is assumed to be controlled by means of a state feedback control action (24) with gain

$$K = \begin{bmatrix} 9.4717 & 0.2980 \\ -3.7750 & 21.5921 \end{bmatrix} \tag{48}$$

The plant is monitored using a state observer (29) with gain

$$L = \begin{bmatrix} 2.8418 & 0.8133 \\ -1.3270 & 5.1801 \end{bmatrix} \tag{49}$$

1) Watermarking signal injection: The procedure developed in Section II is used in order to switch the system set-points. The considered gain matrices for the difference system are

$$M' = \begin{bmatrix} -9.5571 & 0.7519 & 95.4328 & -4.9870 \\ -3.6550 & -23.8981 & -1.0739 & 460.3258 \end{bmatrix}$$
$$F_\Delta = \begin{bmatrix} -0.0150 & -0.0012 \\ -0.0003 & -0.0124 \end{bmatrix} \quad B_s = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{50}$$

and the system set-point is randomly switched between points separated

$$\Delta y_r = [\delta y_1, \delta y_2]^T = [0.5, 0.75]^T \tag{51}$$

around the point $y^{nom} = [3, 2]^T$.

Figure 2 shows the plant outputs response to the injected watermarking signal. It also shows (in green), the system outputs as seen from the controller point of view which are immunized to the injection of the authentication signal.

2) Replay attack: It is assumed that an attacker has hijacked the sensors and secretly records sets of sensor measurements in the stationary. The considered record windows are (in orange in Fig. 2): $\bar{y}_1$ constitutes the record of output 1 from $t_0^1 = 250s$ to $t_f^1 = 350s$ and $\bar{y}_2$ the record of output 2 from $t_0^2 = 500s$ to $t_f^2 = 575s$. At some point in time ($t_{init}^1 = 1150s$, $t_{init}^2 = 1275s$), the attacker replays in loop the previous recording, as shown (also in orange) in Fig. 3. This figure represents how at some random moment, the system set-point switches, causing a mismatch between the injected control action and the system measurements.

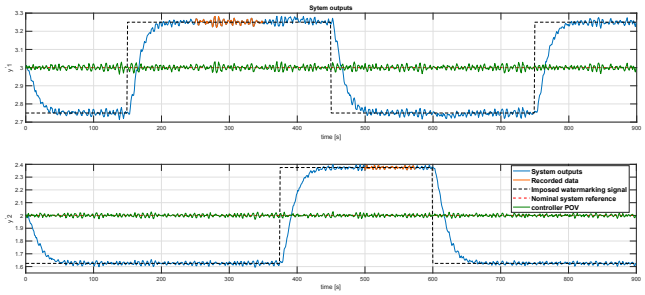

Fig. 2. Watermarking signal injection

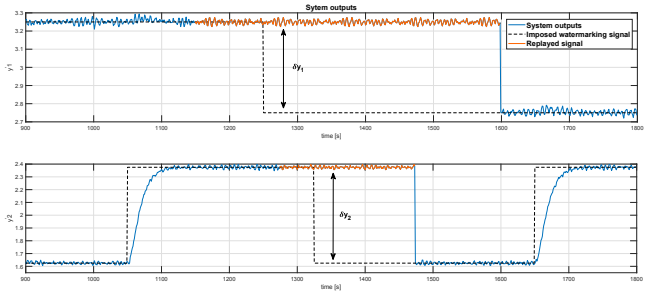

Fig. 3. Replayed measurements

3) Residual analysis: Figure 4 shows the zonotopic representation of the healthy residual RPI set as well as the attack residual RPI sets when a reference mismatch appears in the first output $\Delta y_r = [\delta y_1, 0]^T$, in the second output $\Delta y_r = [0, \delta y_2]^T$ or in both outputs $\Delta y_r = [\delta y_1, \delta y_2]^T$ at the same time.

The computation of these sets is performed offline and provides beforehand information regarding if the introduced $\Delta y_r = [\delta y_1, \delta y_2]^T$ assures attack detection (attack sets not intersecting the healthy set), and attack isolation (attack sets neither intersecting among them). Note that for the selected $\Delta y_r = [0.5, 0.75]^T$ values, the conditions of detectability and isolation are fulfilled.

Related with the temporal attack evolution shown in Figure 3, Figure 4 shows how after a transient stage the computed residuals stabilize within the healthy (blue), first output

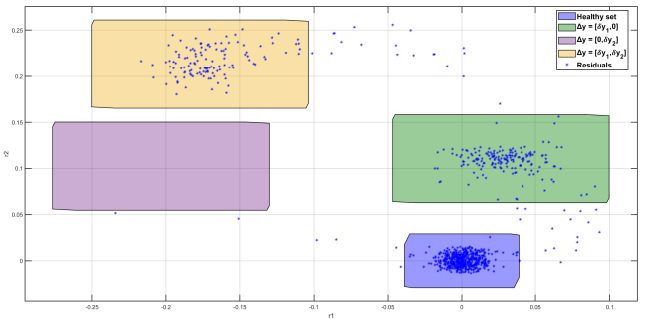

Fig. 4. Healthy/attacked residual sets

(green) and both outputs (yellow) attacked sets, respectively. No residuals stabilize within the second output attacked set, as the second system output never is attacked alone.

V. Conclusions

This work has introduced a novel methodology for injecting a watermarking signal without affecting the predefined system controller, by exploiting the well-known VA scheme. Besides, a set-based analysis concerning the replay attack detectability as a function of the performance loss imposed by an authentication signal, is developed for the case that a state observer is used in order to monitor the plant. The computation of the optimal observer gain such that maximizes the attack detectability while minimizing the required performance loss is a future research direction.

References

[1] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems." in *HotSec*, 2008.

[2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[3] J. P. Conti, "The day the samba stopped [power blackouts]," *Engineering & Technology*, vol. 5, no. 4, pp. 46–47, 2010.

[4] S. Amin, A. A. Cárdenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *International Workshop on Hybrid Systems: Computation and Control*. Springer, 2009, pp. 31–45.

[5] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[6] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2009, pp. 911–918.

[7] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, and J. Schröder, *Diagnosis and fault-tolerant control*. Springer, 2006, vol. 2.

[8] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "A new watermarking approach for replay attack detection in lqg systems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 5143–5148.

[9] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Cost-effective watermark based detector for replay attacks on cyber-physical systems," in *2017 11th Asian Control Conference (ASCC)*. IEEE, 2017, pp. 940–945.

[10] H. S. Sánchez, D. Rotondo, T. Escobet, V. Puig, J. Saludes, and J. Quevedo, "Detection of replay attacks in cyber-physical systems using a frequency-based signature," *Journal of the Franklin Institute*, 2019.

[11] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "Replay attack detection in a multi agent system using stability analysis and loss effective watermarking," in *2017 American Control Conference (ACC)*. IEEE, 2017, pp. 4778–4783.

[12] B. Tang, L. D. Alvergue, and G. Gu, "Secure networked control systems against replay attacks without injecting authentication noise," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 6028–6033.

[13] J. Stoustrup, "Plug & play control: Control technology towards new challenges," *European Journal of Control*, vol. 15, no. 3-4, pp. 311–330, 2009.

[14] F. Stoican, "Fault tolerant control based on set-theoretic methods." Ph.D. dissertation, Supélec, 2011.

[15] V. T. H. Le, C. Stoica, T. Alamo, E. F. Camacho, and D. Dumur, *Zonotopes: From guaranteed state-estimation to control*. John Wiley & Sons, 2013.

[16] M. Pourasghar, V. Puig, and C. Ocampo-Martinez, "Interval observer-based fault detectability analysis using mixed set-invariance theory and sensitivity analysis approach," *International Journal of Systems Science*, pp. 1–22, 2019.

[17] E. Kofman, H. Haimovich, and M. M. Seron, "A systematic method to obtain ultimate bounds for perturbed systems," *International Journal of Control*, vol. 80, no. 2, pp. 167–178, 2007.